***(1 a) [5] How does Flume achieve 100% reliability to the data flow? Explain in brief.***

*Flume uses* ***Transaction*** *in channel, which means events are deleted from channel, only when* ***sinks successfully delivered that event to its destination****.*

*(1 b) [3] Following lines are some of the sample lines taken from a Flume configuration file which has a*

*major mistake in it. Find out the mistake and explain why it's a mistake.*

*agent1. sources = source1*

*agent1. channels = ch1 ch2*

*agent1.sinks = sink1*

*agent1.sources, source1.channels= ch1 ch2*

*agent1, sources, source1, type= spooldir*

*agent1. Sources. sourcel. spoolDir=/tmp/spooldir*

*agent1.sinks.sink1.type = logger*

*agent1. sinks.sink1.channel=ch1 ch2*

*agent1.channels.ch1.type = file*

*agent1. channels. ch2. type = memory*

*agent1, channels. ch2, capacity= 200*

*answer*

*1.b A source instance can specify/attached multiple channels, but a sink instance* ***can only be attached one channel****.*

*(2 a) (4] Why is it said that sharding (horizontal partitioning) of a table is automatic in HBase?*

*Answer*

As a distribution unit in HBase, Regions cause automatic horizontal partitioning. Region Splits happening automatically *when the region grows and crosses the configurable size* ***threshold(256mb***

*(2 b) [4] How does HBase achieve random reads with millisecond latency even though its storage level is HDFS which is not good for random reads?*

*answer.*

*2(b)*

*There are one* ***BlockCache inside each Region Server****, which mean data from all regions of RS share the same cache pool.* ***Client saves the RegionServer info in its own cache*** *which saves time for future reads and there's* ***no need to go again to Zookeeper and to hbase:meta table****. And* ***Hbase servicing read block data from BlockCach****. so, can be described as HBase is requesting exact file from HDFS*

*(2 c) [4] Write down the two most important differences between Minor and Major compactions in HBase*

*answer*

*2.c*

*Minor compaction*

*HBase takes* ***smaller*** *HFiles and recommits them to* ***bigger*** *HFiles .*

*Merges store files of* ***a single store*** *inside* ***a single region.***

*Major compaction*

***Merges and rewrites all the HFiles*** *in different regions to one HFile* ***per CF*** *and in the process,* ***drops deleted or expire cells****.*

***Process delete markers, max versions*** *etc. while minor compaction don’t.*

*--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------*

(*3 a) (2] In Spark-Shell, which contexts are available by default? Just write their names*.

answer

3a. **sparkcontext** sc and **sqlconext**

=-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

(3 b) [3] What are the duties and responsibilities of Driver in Spark?

answer

3b.

**Requesting memory and CPU resources** from cluster managers

**Breaking application logic** into stages and tasks

**Sending tasks** to executors and **collecting the results**

*(3 c) [5] Apache Spark revolves around the concept of a Resilient Distributed Dataset (RDD) which is a fault-tolerant collection of elements that can be operated on in parallel. Hadoop uses replication to achieve fault tolerance. How is fault tolerance achieved in Apache Spark RDDs?*

Answer

3.c

Spark's fault tolerance by using the concept of RDD lineage where the information of an **RDD "came to be" is tracked rather than the data itself**. And fault tolerance is **achieved by healing RDDs**.

*Part 4 of 6- Question 4*

(4) [9] Read the given information and then answer the following (a)¿(b) and (c) questions.

Congratulations! You are just hired by a multinational shopping company and they have recently started working on Hadoop. They built a 1000 nodes Hadoop cluster and they want to copy their transactional data which is in MySQL to HDFS for some analysis. The problem is they don't know how to do this! You come for their rescue and say that you can do it! You get the information about that MySQL table which needs to be copied to HDFS. That table sample looks like this:

EmpSalesByCountry table

Country | LastName |FirstName| ShippedDate | OrderID | SaleAmount

USA | Callahan | Laura |1997-01-16 |10380 |1420

USA | Fuller |Andrew |1997-01-01 |10392 |1440

USA | Davolio |Nancy |1997-01-03 |10393 |33021

The other information about this "EmpSalesByCountry" **table** is as follows:

**Connection** String: jdbc:mysql://156.32.65.83:3306

**Username** and Password both are "root"

**Database** name: orderDB

Primary Key is not defined for this table.

Based on this information, answer the following questions:

(*4 a) [5] Which tool would you use for transferring this data to HDFS? Write the command that you'll use*

*to complete this requirement. You are free to choose the destination path in HDFS. Remember that this*

*MySQL table does not declare any PK and no information has been given to you as to how to split up this table,*

answer

4a.

Sqoop import

--**connect** jdbc:mysql://156.32.65.83:3306/**orderDB**

--**username** root -P

--**table** EmpSalesByCountry

--fields-terminated-By ‘\t’

--**target-dir** =/user/cloudera/sqoopImported

--m 1

(*4 b) (3] Now you are ready with the command that you created for question (4 a) but suddenly you come to know that there's a change in the requirements. Instead of transferring the whole data from this table, only those records should be brought in HDFS where Country is "USA". Your boss also informs you that there will be a need to periodically send these new records of "USA" to HDFS. So how would you modify your command of Q4a to support these new requirements? Write the new command.*

*(You can assume that the original "EmpSalesByCountry» table qets truncated (emptied) after each of these transfers.)*

answer

4b .

Sqoop import

--connect jdbc:mysql://156.32.65.83:3306/orderDB

--username root -P

--table EmpSalesByCountry

--**where** “country = ‘USA’”

--fields-terminated-By ‘\t’

--target-dir =/user/cloudera/sqoopImported

--m 1

--**append**

4 *c) [1] what will be the name of the file in HDFS (after the transfer is done)?*

Answer; 4c **part-m-00000**

*(Q 5) (6] Consider the same transferred file (from Q4 above) which is now stored in HDFS and has grown to Be a little over 2PB. One fine morning when you are running some MR job on that file, one of your colleagues, who is a business analyst (BA), comes to you and says that he wants to execute SQL queries to do some analysis on that same file. Obviously, he cannot run SQL directly on top of that HDFS file, so he is asking you for some help. (Wow...you are becoming the "go-to-quy') What would you do now so that the BA can execute his SQL queries on that file?*

*Find out the schema from the given sample data in Q4 and write down all the required commands properly,*

answer

5 . Since 2PB of data cannot be loaded into a db, Sqoop will not be useful for this need unless we can narrow down the required data scope significantly. Therefore, Hive is ideal for this purpose. Looking at what queries and data he plans to run, I'd create a Hive table or groups on tables on the available data. Then Hive queries can be used to query the data

**CREATE EXTERNAL TABLE** EmpSalesByCountry (

country STRING,

lastName STRING,

firstName STRING,

shippedDate DATE,

orderID INT,

saleAmount INT)

**ROW FORMAT DELIMETED**

**FIELDS** TERMINATED BY ','

**LINES** TERMINATED BY '\n'

**LOCATION** '/user/cloudera/Sura';

(6*) [15] Answer the following multiple-choice questions. Select the answer that you thin one.*

Q*uestion 1 of 14*

*HBase is \_\_\_, need to define only table name and column families at the table*

A. row oriented

**B. schema-less**

C. append only

D. fixed schema db

*Question 2 of 14*

*Mainly what type of data does Flume ingest in Hadoop?*

**A. Streaming**

B. Batch

C. Any type of data

D. None of them!

*Question 3 of 14*

*State True/False: A Region in HBase contains multiple Stores, one for each CF.*

Answer; A Region contains multiple Stores, one for each Column Family -->TRUE

*Question 4 of 14*

*In Apache Hive, on dropping a managed table*

A. The schema gets dropped without dropping the data

B. The data gets dropped without dropping the schema

C. An error is thrown

**D. Both the schema and the data are dropped.**

*Question 5 of 14*

*In HBase,\_\_\_\_\_\_command fetches the contents of a row or a cell.*

A. SELECT

**B. GET**

C PUT

D. SCAN

*Question 6 of 14*

*State True/False: If we launch a Spark application using Spark submit utility, application jar does not*

*get distributed to all worker nodes.*

**False**

*Question 7 of 14*

*From Base shell, a table can be*

A. dropped directly

**B. dropped after disabling.**

C. only disabled, not dropped.

D. only compressed, not dropped.

*Question 8 of 14*

*The partitioning of a table in Hive creates more*

**A. Subdirectories under the table name**

B. Files under the table name

C. Files under database name

D. Subdirectories under the database name

*Question 9 of 14*

*State True/False : In Hive, ORDER BY clause is much faster than SORT BY.*

**False**

*Question 10 of 14*

*How many MapReduce jobs will be submitted by Sqoop while copying an RDBMS table into HDFS?*

A. 1

B. 2

C. 4

**D. Depends upon Sqoop configuration file**

E. Can't tell, it'll be decided by Sqoop during run time.

*Question 11 of 14*

*State True/False: It is possible to create multiple tables in Hive over the same data set.*

**True**

*Question 12 of 14*

*State True/False: In Apache Spark, persist () allows the user to specify the storage level whereas.*

*cache) uses the default storage level.*

**True**

*Question 13 of 14*

*In Sqoop, if you need to import tables with maximum possible speed, which one of the following import*

*control arguments will you use?*

**A.--direct**

B--dump

C. --fastload

D.—bulkload

*Question 14 of 14*

*Given an instance of SparkContext named sc, review the following code*

List<Integer> list = Arrays.asList(1,**30**, **40**, 23, 2, 3, **10**,**46**);

JavaRDD<Integer> numbers = sc.parallelize(list);

JavaPairRDD<String, Integer> numLen = numbers.filter(t->t>= 10&&(t%2==0))

.mapToPair(n -> new Tuple2<String, integer>(“total Even numbers =”, n))

.reduceByKey((x, y) -> x + y);

numLen,saveAsTextFile("output");

*Which one of the following outputs will be produced by the above code?*

A. (Total Even numbers = 5)

**B. (Total Even numbers = 126)**

C. (Total Even numbers = 4)

D. (Total Even numbers = 128)

E. Program will not run; error in the code.